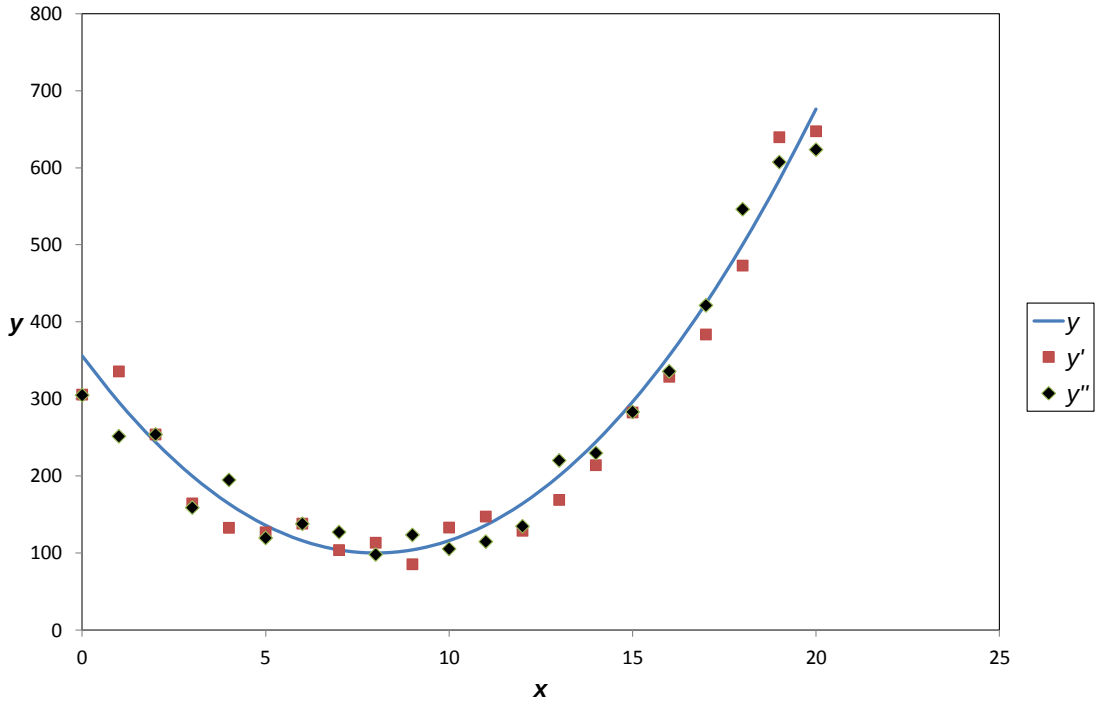
Session A: Least Squares Fitting

A.1 Introduction

Suppose we have some experimental data that should follow the equation:

$$y = 4 \times (x - 8)^2 + 100$$

Of course there may be some experimental error associated with the recording of any particular value. Thus if we take the data on more than one occasion we shall obtain slightly different results. In the figure below the "true" values are indicated by the line y, and the two different sets of measurements at integer values of x are indicated as points in the series y' and y''.



Now the question arises, given the experimental values only, what is the best estimate of the unknown line y vs x? This is where *least squares fitting* comes in. We shall start with *unweighted* least squares – the addition of *weighting* will come later.

Say we have a table of n pairs of measurements, x and y. We wish to find the best fit for a function of y expressed as a function of x; the *regression* of y on x. It is conventional to minimize the sum of squares of the deviations of the data values from the fitted line – hence the term "least squares".

The sum of the squares of the deviations, S, is given by:

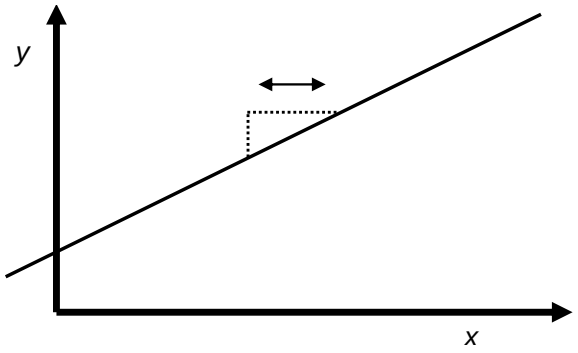
$$S = \sum_{i=1}^{n} [y_i - y_{\text{fit}}(x_i)]^2$$

and we seek the minimum value of S. y_i are the measured values of y, and y_{fit} are the values of y from the expression linking x and y evaluated at the measured x_i. (Note that we could seek the

relation of x expressed as a function of y, which is the regression of x on y. In which case we would minimise a different function, S':

$$S' = \sum_{i=1}^{n} [x_i - x_{\text{fit}}(y_i)]^2$$

where x_{fit} are the values of x from the expression linking y and x, evaluated at the measured y_i. In general the two regressions, y on x and x on y do not produce the same answers. In the special case of a straight line, however they do because the deviations in the y direction are then proportional to the deviations in the x direction linked via the gradient of the line.)

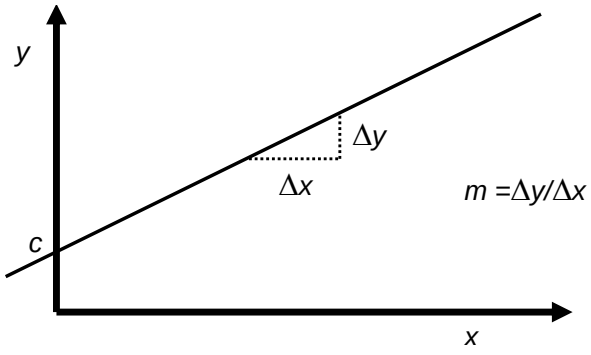


A.2 Straight line fitting

Let us consider the case of the regression of y on x for a straight line. This is often expressed as:

$$y = mx + c$$

where m is the gradient, and c is the intercept.



For consistency with later analysis we shall use the following notation:

$$y = p_1 x + p_0$$

In this way we can build up to higher order equations in a natural way. For example, a 2^{nd} order equation may be written as:

$$y = p_2 x^2 + p_1 x + p_0$$

and hence n^{th} order would be of the form:

$$y = p_n x^n + p_{n-1} x^{n-1} + \cdots + p_1 x + p_0$$

But to return to 1st order (straight line fitting), we need to minimize:

$$S = \sum_{i=1}^{n} [y_i - (p_1 x_i + p_0)]^2$$

where we have substituted the expected relation "$p_1 x_i + p_0$" for y_{fit}. We need to minimize S with respect to the two parameters, p_1 and p_0. Obviously we differentiate S with respect to both and set the results to zero:

$$\frac{\partial S}{\partial p_1} = 2 \sum_{i=1}^{n} [y_i - (p_1 x_i + p_0)] (-x_i) = 0$$
$$\frac{\partial S}{\partial p_0} = 2 \sum_{i=1}^{n} [y_i - (p_1 x_i + p_0)] (-1) = 0$$

These can be rearranged as:

$$p_1 \sum_{i=1}^{n} x_i^2 + p_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

and

$$p_1 \sum_{i=1}^{n} x_i + p_0 \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i$$

respectively. In fact, the last of these can be written as:

$$p_1 \sum_{i=1}^{n} x_i + n p_0 = \sum_{i=1}^{n} y_i$$

We can thus solve the two equations for the two unknowns, p_1 and p_0 and obtain the standard least squares straight-line fitting equations:

$$p_1 = \frac{\left[n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right]}{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right]} \qquad p_0 = \frac{\left[\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \right]}{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right]}$$

Exercise A.1

1. Take the following table of values of x and y and calculate the straight line of best fit. That is, evaluate the values of p_1 and p_0 using the equations above.

x	y
0	0.97
0.5	2.51
1.0	4.09
2.0	6.92
3.5	11.52
4.0	13.01
5.0	16.02

2. Check your results using Excel's trendline (insert a scatter chart of y vs x, right click a data point on the chart and select "Add Trendline..." - make sure "Display Equation on chart" is selected so that you can see the parameter values!). Alternatively you can use the program POLYFIT. See the document *Using the Computing Facilities* in the *LM PH605 Practical Skills for Reactor Physics* section of Canvas for instructions on how to do this.
3. Now do the regression of x on y. Do you get the expected answers?
4. Go back to the regression of y on x. Using the spreadsheet multiply all the x values by 10, then redo the calculation (do this by multiplying the x values by the content of another cell). Repeat for higher and higher multiples of 10. Do you get the result you expect for 10^{10}, 10^{100}, 10^{200}, etc? If not, why not?

A.3 Higher order fitting: second order polynomials.

Consider for a moment the equations for the straight line fit:

$$p_1 \sum_{i=1}^{n} x_i^2 + p_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

and

$$p_1 \sum_{i=1}^{n} x_i + p_0 \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i$$

If we note that $x_i^0 = 1$ then we can rewrite the second equation as:

$$p_1 \sum_{i=1}^{n} x_i + p_0 \sum_{i=1}^{n} x_i^0 = \sum_{i=1}^{n} x_i^0 y_i$$

which will help us see the symmetry in the equations.

Let us use "Gaussian notation", where $[x^m]$ implies $\sum_{i=1}^{n} x_i^m$; hence $\left[x^0\right] = \sum_{i=1}^{n} x_i^0 = n$.

We can then write the equations in a compact form:

$$p_1[x^2] + p_0[x^1] = [x^1 y]$$
$$p_1[x^1] + p_0[x^0] = [x^0 y]$$

This can be written in matrix form as:

$$\begin{pmatrix} [x^2] & [x^1] \\ [x^1] & [x^0] \end{pmatrix} \begin{pmatrix} p_1 \\ p_0 \end{pmatrix} = \begin{pmatrix} [x^1 y] \\ [x^0 y] \end{pmatrix}$$

or

$$A \begin{pmatrix} p_1 \\ p_0 \end{pmatrix} = \begin{pmatrix} [x^1 y] \\ [x^0 y] \end{pmatrix} \qquad \text{where} \qquad A = \begin{pmatrix} [x^2] & [x^1] \\ [x^1] & [x^0] \end{pmatrix}.$$

The solution for the parameters p_1 and p_0 are then obtained by inverting the matrix A:

$$\begin{pmatrix} p_1 \\ p_0 \end{pmatrix} = A^{-1} \begin{pmatrix} [x^1 y] \\ [x^0 y] \end{pmatrix}$$

The second order polynomial least squares fit is obtained in an exactly equivalent manner. The symmetry of the equations in matrix form goes through an obvious progression:

$$A \begin{pmatrix} p_2 \\ p_1 \\ p_0 \end{pmatrix} = \begin{pmatrix} [x^2 y] \\ [x^1 y] \\ [x^0 y] \end{pmatrix} \qquad \text{where} \qquad A = \begin{pmatrix} [x^4] & [x^3] & [x^2] \\ [x^3] & [x^2] & [x^1] \\ [x^2] & [x^1] & [x^0] \end{pmatrix}$$

The solution is then:

$$\begin{pmatrix} p_2 \\ p_1 \\ p_0 \end{pmatrix} = A^{-1} \begin{pmatrix} [x^2 y] \\ [x^1 y] \\ [x^0 y] \end{pmatrix}$$

Exercise A.2

1. Take the following tables of pairs of values of x and y and calculate the best fit of a second order polynomial, i.e. evaluate the values of p_2, p_1 and p_0 using the equations above. There are two datasets generated with random differences in the y values.

x	y
0	305.5392
1	335.6331
2	253.8043
3	164.2771
4	132.6123
5	126.8101
6	138.0547
7	103.7072
8	113.2603
9	85.29682
10	132.9629
11	147.2522
12	128.7884
13	168.8312
14	213.9263
15	282.3310
16	328.6762
17	383.6006
18	473.0220
19	639.6690
20	647.3819

x	y
0	304.8299
1	251.4010
2	253.8037
3	158.8323
4	194.7284
5	119.4289
6	137.8466
7	127.0196
8	97.92346
9	123.4052
10	105.3041
11	114.6574
12	134.6610
13	220.0674
14	229.6439
15	283.0500
16	335.7672
17	421.3725
18	546.1925
19	607.3927
20	623.5712

2. Compare your results with those obtained using Excel's Trendline (polynomial fit) or POLYFIT.

3. Given that the original function from which the values in the table were obtained is:

$$y = 4 \times (x - 8)^2 + 100$$

are the parameters of the polynomial as expected?

A.4 Higher order fitting

The extension to higher order polynomials should be obvious. The matrix A grows in a symmetrical way; if the polynomial order is n, then the top left hand element of A is $[x^{2n}]$, and there are successive diagonals all with successively lower order power terms until the bottom right element of A is $[x^0]$ as before.

Exercise A.3

1. Take the tables from Exercise A.2 and calculate the best fit of a 3^{rd} order polynomial, i.e. evaluate the values of p_3, p_2, p_1 and p_0 using Excel or POLYFIT.
2. Now generate the values for the expected polynomial $y = 4 \times (x - 8)^2 + 100$. Use Excel or POLYFIT again and fit to a second and then third order polynomial. Are the p_3 values from this and the preceding polynomial fits as you might expect?

A.5 Lower order fitting

The logic of the progression for the first to second to third and higher order polynomials can in fact be taken in the reverse direction. A *zero* order fit can be made.

Exercise A.4

Demonstrate that the zero order fit is equivalent to taking the average of the y values.

A.6 Residuals

We might not always know what order of polynomial is appropriate to fit to our data. For example, when we calibrate a γ-ray detector, we measure the channel numbers, k, at which known γ-ray energies, E, appear. This allows us to convert from measured channel numbers to energies, which then allows us to identify unknown isotopes from their measured γ spectra. It is possible that there

is a small second order contribution to the calibration, i.e. a term that depends on E^2, which would give a more accurate measure for the energy compared with a first order (linear) fit. We can get an idea if this is the case by examining the *residuals* of the fit.

The residuals are the differences between the measured values, y_i, and the values calculated using the fit, $y_i(\text{fit}) = p_0 + p_1 x_i (+p_2 x_i^2 + ...)$:

$$r_i = y_i - y_i(\text{fit}) \tag{A.1}$$

Plotting the residuals against x can reveal if there is a systematic variation from the fit line, which could be removed by including higher order terms in the polynomial. If the fit is appropriate, the residuals will be randomly scattered about the x-axis, due to the uncertainties inherent in every measurement (see next section).

It must be noted, however, that it is not appropriate to arbitrarily add extra terms to improve the fit if there is a known theoretical relationship between x and y. The aim is to be able to explain your data, and extract useful information from it, not to make S as small as you can by adding in more and more parameters that are meaningless! You can always make $S = 0$ if you have as many parameters as data points (you can always draw a straight line between two points: 2 data points, 2 parameters), but this doesn't tell us anything useful.

Exercise A.5

1. Take the following table of pairs of values of x and y and calculate the values of p_1 and p_0 for the best fit line. Calculate the residuals for each point and plot them against x.

x	y
0	4.492
1	7.195
2	9.894
3	12.616
4	15.275
5	17.948
6	20.609
7	23.276
8	25.927
9	28.592
10	31.198
11	33.836
12	36.456
13	39.105
14	41.715
15	44.335
16	46.928
17	49.519
18	52.131
19	54.723
20	57.333

2. Now fit a second order polynomial to the data, and plot the residuals for this fit.
